Computational Statistics and Data Analysis 55 (2011) 12–25

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Alternating imputation posterior estimation of models with crossed random effects

S.-J. Cho [a,*], S. Rabe-Hesketh [b,c]

[a] Peabody College, Vanderbilt University, Peabody Hobbs 213A, Nashville TN 37203, United States
[b] Graduate School of Education, University of California, Berkeley, United States
[c] Institute of Education, University of London, UK

ARTICLE INFO

Article history:
Received 16 March 2009
Received in revised form 16 April 2010
Accepted 16 April 2010
Available online 27 April 2010

Keywords:
Adaptive quadrature
Crossed random effects
Generalized linear mixed model
Item response theory
Laplace approximation
Random cross-classification
Salamander mating data
Two-way error components

ABSTRACT

Generalized linear mixed models or latent variable models for categorical data are difficult to estimate if the random effects or latent variables vary at non-nested levels, such as persons and test items. Clayton and Rasbash (1999) suggested an Alternating Imputation Posterior (AIP) algorithm for approximate maximum likelihood estimation. For item response models with random item effects, the algorithm iterates between an item wing in which the item mean and variance are estimated for given person effects and a person wing in which the person mean and variance are estimated for given item effects. The person effects used for the item wing are sampled from the conditional posterior distribution estimated in the person wing and vice versa. Clayton and Rasbash (1999) used marginal quasi-likelihood (MQL) and penalized quasi-likelihood (PQL) estimation within the AIP algorithm, but this method has been shown to produce biased estimates in many situations, so we use maximum likelihood estimation with adaptive quadrature. We apply the proposed algorithm to the famous salamander mating data, comparing the estimates with many other methods, and to an educational testing dataset. We also present a simulation study to assess performance of the AIP algorithm and the Laplace approximation with different numbers of items and persons and a range of item and person variances.

1. Introduction

Models with crossed random effects arise when units are nested in several types of clusters, or classifications, that are not nested. A classical example in education is students nested in secondary schools cross-classified by primary schools (e.g., Goldstein, 2003, p. 191–192) or students nested in schools cross-classified by neighborhoods (e.g., Raudenbush and Bryk, 2002, Chapter 12). Another example from education is where students are nested in teachers in a given year, but the teachers can change over time (Raudenbush, 1993; McCaffrey et al., 2004). In economics, a canonical example is panel data where both units and panel waves can be viewed as clusters, giving rise to two-way error components models (e.g., Baltagi, 2001, pp. 33–39). In psychometrics, item response models for educational assessment data sometimes include latent variables for persons and test items (e.g., Van den Noortgate et al., 2003). See Browne et al. (2001) for other examples.

Consider a simple example of an item response model with a random item parameter

$$\text{logit}\left[\Pr(y_{ij}=1|\zeta_{1j},\zeta_{2i})\right]=\beta_1+\zeta_{1j}+\zeta_{2i}, \tag{1}$$

where β_1 is the average logit of the probability of a correct response, averaging over persons ($j=1,\ldots,J$) and items ($i=1,\ldots,I$), $\zeta_{1j}\sim N(0,\psi_1)$ is a random person ability parameter, and $\zeta_{2i}\sim N(0,\psi_2)$ is a random item difficulty

* Corresponding author.
E-mail addresses: sj.cho@vanderbilt.edu (S.-J. Cho), sophiarh@berkeley.edu (S. Rabe-Hesketh).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.04.015

parameter. Here the random effects or latent variables ζ_{1j} and ζ_{2i} are crossed since every item is offered to all persons and every person responds to all items.

Maximum likelihood estimation of models for categorical data with crossed random effects is challenging. This is because the marginal likelihood does not have a closed form so that maximum likelihood estimation requires numerical or Monte Carlo integration. If the random effects are nested, the integrals are also nested (e.g., Rabe-Hesketh et al., 2005), keeping the computational burden low. Models with crossed random effects can be reformulated as models with nested random effects (Goldstein, 1987; Rasbash and Goldstein, 1994), but this approach requires evaluation of high-dimensional integrals and is therefore computationally demanding. Specifically, an equivalent model to Eq. (1) can be written as

$$\text{logit}\left[\Pr(y_{ij}=1|\zeta_{1j}^{(2)},\zeta_{2i}^{(3)})\right]=\beta_1+\zeta_{1j}^{(2)}+\sum_{a=1}^{I}\zeta_{2a}^{(3)}X_{ai}, \tag{2}$$

where the superscripts denote the levels of the model, level 2 being the person and level 3 the entire dataset. The variable $X_{ai}=1$ if $a=i$ and $X_{ai}=0$ if $a\neq i$, so that the last term evaluates to $\zeta_{2i}^{(3)}$ as required (see also Rabe-Hesketh and Skrondal, 2008, pp. 489–493). Unfortunately, this approach involves an I-dimensional integral at level 3 and is therefore computationally demanding unless the number of items is small.

Several closely related approximations to maximum likelihood estimation have been proposed to avoid high-dimensional numerical integration in generalized linear mixed models, including marginal quasi-likelihood (MQL, Goldstein, 1991), penalized quasi-likelihood (PQL, Breslow and Clayton, 1993) and its second-order improvement (PQL-2, Goldstein and Rasbash, 1996), bias-corrected PQL (Breslow and Lin, 1995; Lin and Breslow, 1996), Laplace approximations (Tierney and Kadane, 1986; Pinheiro and Bates, 1995; Raudenbush et al., 2000), and the hierarchical-likelihood method (Lee and Nelder, 1996, 2006). However, both MQL and PQL perform poorly for dichotomous response with small cluster sizes, with a downward bias in the estimated variance components (Rodríguez and Goldman, 1995; Goldstein and Rasbash, 1996; Raudenbush et al., 2000). Although PQL-2 performs considerably better than PQL, this downward bias often remains a problem (Rodríguez and Goldman, 2001; Breslow, 2004; Browne and Draper, 2006). Joe (2008) found similar results for the Laplace approximation for binary responses with small cluster sizes. Diaz (2007) found that the higher-order Laplace approximation proposed by Raudenbush et al. (2000) reduces the bias of PQL but increases the mean squared error. Surprisingly little work has been done on evaluating these and other approximate methods.

Bellio and Varin (2005) proposed a pairwise likelihood approach in which the product of pairwise marginal likelihoods is maximized. Tibaldi et al. (2007) developed a conditional mixed model approach combined with pseudolikelihood estimation. Specifically, they treat all possible item pairs as matched pairs in conditional logistic regression to eliminate the person random effect and estimate the item variance. They then repeat the procedure with the role of items and persons reversed to estimate the person variance. The method works only if there is no more than one observation for each combination of the levels of the crossed random effects.

Perhaps the most straightforward approach to estimation of models with crossed random effects is Markov chain Monte Carlo (MCMC, Karim and Zeger, 1992; Rasbash and Browne, 2007). For the Gibbs sampler, the conditional posterior distributions remain the same as for models with fixed item parameters, except that the conditional distributions for the item parameters now depend on the hyperprior for the item variance. However, MCMC is computationally expensive, and it may be difficult to specify vague hyperpriors for the variance parameters in hierarchical models that result in a posterior mean (or mode) close to the maximum likelihood estimate (Natarajan and Kass, 2000; Browne and Draper, 2006). When the number of higher-level units is small, the choice of prior distribution becomes more important (Lambert, 2006). Gelman (2006) suggested using the half-t or half-Cauchy distribution in this case.

McCulloch (1994) applied a Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990) for maximum likelihood estimation of generalized linear mixed models with crossed random effects. In the E-step, a Gibbs sampler is used to sample the random effects from their posterior distributions and the M-step is the estimation of a generalized linear mixed model. Booth and Hobert (1999) implemented MCEM employing importance sampling in the E-step and Vaida and Meng (2005) employing a slice sampler to fit generalized linear models with crossed random effects. This algorithm is also computationally expensive requiring many draws of the random effects in the E-step to achieve a sufficiently small Monte Carlo error close to convergence.

Similar to MCMC and MCEM, the alternating imputation-posterior algorithm (AIP, Clayton and Rasbash, 1999) makes use of data augmentation. The aim is to obtain approximate maximum likelihood estimates. The algorithm alternates between an item wing in which item difficulties are sampled for given person abilities and a person wing in which person abilities are sampled for given item difficulties, by sampling from the respective conditional posterior distributions. However, instead of drawing individual model parameters from their posterior distributions as in MCMC, β_1 and $\log(\psi_1)$ are estimated by maximum likelihood in the person wing (for given item parameters), and then sampled from their estimated sampling distribution. Similarly, β_1 and $\log(\psi_2)$ are estimated by maximum likelihood in the item wing (for given person parameters). Each maximum likelihood estimation involves only one random effect and can therefore be accomplished relatively easily.

Unlike MCMC, the AIP algorithm does not require specification of prior distributions for the model parameters. Furthermore, the algorithm typically converges much more rapidly because several model parameters are updated simultaneously. As pointed out by Clayton and Rasbash (1999), following an initial "burn-in", this algorithm also requires many fewer draws than a Gibbs sampler based on scalar nodes to estimate posteriors accurately. The reason for this is that

characteristics of the joint posterior distribution can be estimated using Rao–Blackwellization (Gelfand and Smith, 1990). While it would generally be possible to use Rao–Blackwellization within MCMC, it is not usually done.

Clayton and Rasbash (1999) used MQL and PQL within the AIP algorithm, as implemented in MLwiN (Goldstein et al., 1998). As discussed above, MQL and PQL sometimes underestimate the variance components. Clayton and Rasbash (1999) found that this problem can also occur when MQL and PQL are used within their AIP algorithm. For a simulated dataset, AIP with PQL-2 produced a variance estimate that was just over half the true value or MCMC estimate. We therefore develop an AIP algorithm that uses maximum likelihood estimation with adaptive quadrature (Pinheiro and Bates, 1995; Bock and Schilling, 1997; Rabe-Hesketh et al., 2005). Adaptive quadrature is an improvement of Gauss–Hermite quadrature (Bock and Lieberman, 1970; Butler and Moffitt, 1982; Hedeker and Gibbons, 1994) that performs well in a wide range of situations (Rabe-Hesketh et al., 2005). For sampling item difficulties and person abilities, both a normal and a discrete approximation are developed for the corresponding posterior distributions.

In Section 2, we describe the AIP algorithm with adaptive quadrature. We then apply the algorithm to the famous salamander mating data and compare the results with estimates using a large range of alternative estimation methods. We also analyze the simulated dataset considered by Clayton and Rasbash (1999) to compare our estimates with theirs. Item response models with random item parameters are discussed in Section 3 and fitted to a dataset using several different methods. In Section 4, we present a simulation study designed to examine the performance of the proposed algorithm compared with the Laplace approximation. We end with a brief discussion in Section 5.

2. Method

2.1. AIP algorithm

Clayton and Rasbash (1999) suggested a special kind of Markov chain Monte Carlo (MCMC) algorithm for generalized linear mixed models with crossed random-effects based on the imputation posterior (IP) algorithm of Tanner and Wong (1987, pp. 90–92) which can be outlined as follows (Tanner, 1996):

I-step (data augmentation): Impute missing data (random effects) by sampling from the distribution of the missing data conditional on the observed data. This requires first sampling the parameters from the current approximation of their posterior distribution.

P-step: Update the approximation of the posterior distribution.

For the item response model with a random item parameter, the algorithm consists of two wings, a person wing and an item wing. In the person wing, the item difficulties are treated as known and in the item wing, the person abilities are treated as known. In the person wing, the parameters β_1 and $\log(\psi_1)$ are estimated (P-step) and the abilities ζ_{1j} sampled (I-step) by first sampling parameters from their approximate posterior distribution (treating the item difficulties as known). In the item wing, the parameters β_1 and $\log(\psi_2)$ are estimated (P-step) and item difficulties ζ_{2i} sampled (I-step), again by first sampling the parameters from their approximate posterior distribution (treating the person abilities as known).

Specifically, after setting initial values $\boldsymbol{\zeta}_2^0$ for the item difficulties, the person wing and item wing outlined below are alternated until convergence. In iteration k:

Person wing

Treat the item difficulties $\boldsymbol{\zeta}_2^{k-1} = (\zeta_{21}^{k-1}, \ldots, \zeta_{2I}^{k-1})'$ from the previous iteration as known:

$$\text{logit}\left[\Pr(y_{ij} = 1|\zeta_{1j}, \zeta_{2i}^{k-1})\right] = \beta_1 + \zeta_{1j} + \zeta_{2i}^{k-1}. \tag{3}$$

Let the parameters be denoted $\boldsymbol{\vartheta}_1 = \{\beta_1, \log(\psi_1)\}$.

1. Obtain maximum likelihood estimates $\widehat{\boldsymbol{\vartheta}}_1^k$ with estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta}_1}^k$
2. Sample parameters $\boldsymbol{\vartheta}_1^k$ from their approximate sampling distribution

$$\boldsymbol{\vartheta}_1^k|\boldsymbol{\zeta}_2^{k-1} \sim N(\widehat{\boldsymbol{\vartheta}}_1^k, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta}_1}^k). \tag{4}$$

3. Sample $\boldsymbol{\zeta}_1^k = (\zeta_{11}^k, \ldots, \zeta_{1J}^k)'$ from its conditional posterior distribution with parameters $\boldsymbol{\vartheta}_1^k$.

Item wing

Treat the person abilities $\boldsymbol{\zeta}_1^k = (\zeta_{11}^k, \ldots, \zeta_{1J}^k)'$ from the person wing as known:

$$\text{logit}\left[\Pr(y_{ij} = 1|\zeta_{2i}, \zeta_{1j}^k)\right] = \beta_1 + \zeta_{2i} + \zeta_{1j}^k. \tag{5}$$

Let the parameters be denoted $\boldsymbol{\vartheta}_2 = \{\beta_1, \log(\psi_2)\}$.

1. Obtain maximum likelihood estimates $\widehat{\boldsymbol{\vartheta}}_2^k$ with estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta}_2}^k$
2. Sample parameters $\boldsymbol{\vartheta}_2^k$ from their approximate sampling distribution

$$\boldsymbol{\vartheta}_2^k|\boldsymbol{\zeta}_1^k \sim N(\widehat{\boldsymbol{\vartheta}}_2^k, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta}_2}^k). \tag{6}$$

3. Sample $\boldsymbol{\zeta}_2^k = (\zeta_{21}^k, \ldots, \zeta_{2I}^k)'$ from its conditional posterior distribution with parameters $\boldsymbol{\vartheta}_2^k$.

After convergence is achieved (burn-in, see Section 2.5), the algorithm is continued for a fixed number of iterations and the parameter estimates are obtained by averaging the estimates obtained after burn-in (see Section 2.6).

In the following three sections, we discuss the implementation of steps 1–3.

2.2. Step 1: maximum likelihood estimation using adaptive quadrature

Treating the item difficulties as known (sampled in the item wing as ζ_{2i}^{k-1}), the likelihood maximized in step 1 of the person wing is the product of contributions from persons j, given by

$$\ell_j^k(\boldsymbol{\vartheta}_1) = \int_{-\infty}^{\infty} \mathrm{P}(\mathbf{y}_j, \zeta_{1j}|\boldsymbol{\zeta}_2^{k-1})\mathrm{d}\zeta_{1j} = \int_{-\infty}^{\infty} g(\zeta_{1j}; 0, \psi_1) \prod_i \mathrm{P}(y_{ij}|\zeta_{1j}, \zeta_{2i}^{k-1}; \beta_1)\mathrm{d}\zeta_{1j}$$
$$= \int_{-\infty}^{\infty} \phi(\nu_j) \prod_i \mathrm{P}(y_{ij}|\sqrt{\psi_1}\nu_j, \zeta_{2i}^{k-1}; \beta_1)\mathrm{d}\nu_j,$$

where $g(\zeta_{1j}; 0, \psi_1)$ is the (prior) person ability density, specified as normal with mean 0 and variance ψ_1 and $\phi(\nu_j)$ is the standard normal density. An analogous likelihood is maximized in the item wing.

The integral cannot be evaluated analytically and Gauss–Hermite quadrature is therefore often used (e.g., Bock and Lieberman, 1970). However, it has been shown that the estimates are biased for large cluster sizes and/or intraclass correlations (Albert and Follmann, 2000; Lesaffre and Spiessens, 2001; Rabe-Hesketh et al., 2005) and adaptive quadrature has been shown to perform considerably better (Rabe-Hesketh et al., 2005).

In ordinary quadrature, the kernel $\phi(\nu_j)$ is essentially replaced by a discrete distribution so that the integral becomes a sum. Adaptive quadrature exploits the fact that the integrand, which is proportional to the conditional posterior density of the random effect (for given parameter values), is often well approximated by a normal distribution (see also Section 2.4.1) with person-specific mean μ_j^k and person-specific variance τ_j^k. The integral can be rewritten as

$$\ell_j^k(\boldsymbol{\vartheta}_1) = \int_{-\infty}^{\infty} g(\nu_j; \mu_j^k, \tau_j^k) \cdot \left[\frac{\phi(\nu_j) \prod_i \mathrm{P}(y_{ij}|\sqrt{\psi_1}\nu_j, \zeta_{2i}^{k-1}; \beta_1)}{g(\nu_j; \mu_j^k, \tau_j^k)} \right] \mathrm{d}\nu_{1j}. \tag{7}$$

Treating $g(\nu_j; \mu_j^k, \tau_j^k)$ as the kernel in the Gaussian quadrature approximation is analogous to treating it as importance density in Monte Carlo integration. Changing the variable of integration from ν_j to $z_j = \frac{(\nu_j - \mu_j^k)}{\tau_j^k}$ and applying the standard quadrature rule yields,

$$\ell_j^k(\boldsymbol{\vartheta}_1) \simeq \sum_{r=1}^{R} \pi_{jr}^k \prod_i \mathrm{P}(y_{ij}|\alpha_{jr}^k\sqrt{\psi_1}, \zeta_{2i}^{k-1}), \tag{8}$$

where

$$\alpha_{jr}^k = \sqrt{\tau_j^k}a_r + \mu_j^k, \quad \text{and} \quad \pi_{jr}^k = \frac{\sqrt{\tau_j^k}\phi(\alpha_{jr}^k)}{\phi(a_r)}p_r, \tag{9}$$

and a_r and p_r are Gauss–Hermite quadrature locations and weights respectively.

We can see from Eq. (9) that adaptive quadrature shifts and scales the quadrature locations to place them under the peak of the integrand. Following Naylor and Smith (1982), we use the estimated posterior moments for μ_j^k and τ_j^k. Rabe-Hesketh et al. (2005) developed this approach for multilevel models with an arbitrary number of levels.

Another version of adaptive quadrature, suggested by Liu and Pierce (1994), uses the estimated posterior mode for μ_j^k and the standard deviation of the normal density matching the curvature of the estimated posterior density at the mode for $\sqrt{\tau_j^k}$. This version has been used for two-level nonlinear mixed models (Pinheiro and Bates, 1995) and for exploratory factor analysis with dichotomous responses (Bock and Schilling, 1997; Schilling and Bock, 2005). Pinheiro and Chao (2006) extended the approach to multilevel generalized linear models with more than two levels, but only for canonical link functions.

In the AIP algorithm, the parameters are estimated using Stata's *xtlogit* command (StataCorp, 2007) which employs a Newton–Raphson algorithm with analytical first and second derivatives to maximize the likelihood. The number of quadrature points required is determined by fitting the person-wing model with item difficulties set to 0 and the item-wing model with person abilities set to 0. The number of quadrature points is increased from 5 in 5 point increments. If the change in maximized log-likelihood associated with an increment is less than 1×10^{-10}, the smaller number of adaptive quadrature points is used.

2.3. Step 2: sampling the model parameters

In step 2, the parameters are sampled from a multivariate normal distribution with mean given by maximum likelihood estimates (MLEs) $\widehat{\boldsymbol{\vartheta}}$ and covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta}}$ by the inverse of the estimated information matrix obtained in step 1.

This distribution approximates the fully Bayesian posterior if uniform priors are specified for all parameters. In this case, the posterior distribution is just the normalized likelihood which is approximated by a multivariate normal distribution. A log transformation of the variance parameters is used to improve the normal approximation. In step 2, we can see clear differences between AIP and Gibbs sampling. First, we do not need to specify prior distributions in AIP. Second, a whole vector of nodes is sampled in AIP while a scalar is sampled in Gibbs sampling.

2.4. Step 3: sampling the random effects from their posterior distribution

In the imputation step, the person abilities and item difficulties are sampled from their respective conditional posterior distributions. For the person abilities, the posterior density is

$$\mathrm{P}(\zeta_{1j}|\mathbf{y}_j, \boldsymbol{\zeta}_2^{k-1}; \beta_1^{k,1}) = \frac{g(\zeta_{1j}; 0, \psi_1^k)\prod_i \mathrm{P}(y_{ij}|\zeta_{1j}, \zeta_{2i}^{k-1}; \beta_1^{k,1})}{\int g(\zeta_{1j}; 0, \psi_1^k)\prod_i \mathrm{P}(y_{ij}|\zeta_{1j}, \zeta_{2i}^{k-1}; \beta_1^{k,1})\mathrm{d}\zeta_{1j}}, \tag{10}$$

where $\beta_1^{k,1}$ is the draw from the person wing (wing 1) in iteration k. In the context of imputing ability scores from item response models in education surveys, such random draws are sometimes referred to as plausible values (Mislevy, 1991; Mislevy et al., 1992).

2.4.1. Normal approximation

According to the Bayesian central limit theorem, posterior distributions approach normality as the sample size (here the number of items or number of persons) tends to infinity (see Chang and Stout, 1993, for the binary case). We therefore approximate the posterior by a normal density with person-specific posterior mean μ_j^k and posterior standard variance τ_j^k for parameters $\boldsymbol{\vartheta}_1^k$,

$$\mathrm{P}(\zeta_{1j}|\mathbf{y}_j; \boldsymbol{\zeta}_2^{k-1}, \beta_1^{k,1}) \simeq g(\zeta_{1j}; \mu_j^k, \tau_j^k). \tag{11}$$

We compute the posterior mean and variance using the program *gllamm* and the corresponding prediction command (Rabe-Hesketh and Skrondal, 2008). This normal approximation ignores any skewness for clusters with large or small cluster totals (Thomas and Gan, 1997), and we therefore also consider a discrete approximation.

2.4.2. Discrete approximation

The discrete approximation is based on adaptive quadrature. The values $\alpha_{jr}^k\sqrt{\psi_1^k}$ $(r = 1, \ldots, R)$ are sampled with probabilities

$$\mathrm{P}_r = \frac{\pi_{jr}^k \prod_i \mathrm{P}\left(y_{ij}\middle|\alpha_{jr}^k\sqrt{\psi_1^k}, \zeta_{2i}^{k-1}; \beta_1^{k,1}\right)}{\sum_{r=1}^R \pi_{jr}^k \prod_i \mathrm{P}\left(y_{ij}\middle|\alpha_{jr}^k\sqrt{\psi_1^k}, \zeta_{2i}^{k-1}; \beta_1^{k,1}\right)}, \tag{12}$$

where α_{jr}^k and π_{jr}^k are the adaptive quadrature locations and weights defined in Eq. (9).

In the context of generating plausible values for the National Assessment of Educational Progress, Thomas and Gan (1997) suggested a sampling importance resampling (SIR) approach. R values of the random effect are first sampled from an importance density g, giving the support points ζ_1^r $(r = 1, \ldots, R)$ of the discrete distribution. Probabilities P_r associated with ζ_1^r are then calculated by normalizing $f(\zeta_1^r)/g(\zeta_1^r)$ to sum to 1, where f is the posterior density. Keeping in mind the importance sampling interpretation of adaptive quadrature, our approach can be considered a deterministic version of SIR, with $g(\zeta_{1j}; \mu_j^k, \tau_j^k)$ as importance density (See Eq. (7)). In this paper we use 50 quadrature points for the discrete approximation.

2.5. Convergence checking

The Gelman and Rubin (1992) method is used for checking convergence of the AIP algorithm. This method relies on running at least two independent chains. Let β_1^{kq} denote the kth sampled value from chain q, with $k = 1, \ldots, n$ and $q = 1, \ldots, m$. We compute two quantities, the between-chain variance, B, and the within-chain variance, W, as follows:

$$B = \frac{n}{m-1}\sum_{q=1}^m (\overline{\beta}_1^{\cdot q} - \overline{\beta}_1^{\cdot\cdot})^2, \tag{13}$$

where $\overline{\beta}_1^{\cdot q} = \frac{1}{n}\sum_{k=1}^n \beta_1^{kq}$, and $\overline{\beta}_1^{\cdot\cdot} = \frac{1}{m}\sum_{q=1}^m \beta_1^{\cdot q}$, and

$$W = \frac{1}{m}\sum_{q=1}^m s_q^2, \tag{14}$$

where $s_q^2 = \frac{1}{n-1}\sum_{k=1}^n (\beta_1^{kq} - \overline{\beta}_1^{\cdot q})^2$.

The marginal posterior variance of the estimand can be estimated as follows:

$$V = \widehat{\text{var}}^{+}(\beta_1|y) = \frac{n-1}{n}W + \frac{1}{n}B. \tag{15}$$

Convergence can be monitored by calculating the following potential scale reduction:

$$\sqrt{\widehat{R}} = \sqrt{\frac{\frac{n-1}{n}W + \frac{1}{n}B}{W}}. \tag{16}$$

If $\sqrt{\hat{R}}$ is "near" 1, convergence is achieved.

For parameters that are sampled in both wings (here β_1), Clayton and Rasbash (1999) suggest using the wings as chains. To obtain different chains for parameters sampled in only one of the wings (here $\log(\psi_1)$ and $\log(\psi_2)$), we use different realizations from $\zeta_{2i}^0 \sim \text{N}(0, 2^2)$ as starting values for the item difficulties. An alternative method for generating different chains would be to alternate the order of the wings, i.e., for one chain start with the person wing and for the other chain start with the item wing.

For each parameter common to both wings, there are four pairs of chains that can be used for convergence checking, the two wings for each set of starting values and the two sets of starting values for each wing. For each parameter unique to a wing there is one pair of chains.

Convergence is assessed by calculating sequences of statistics $V(h)$, $W(h)$, and $\widehat{R}(h)$, $h = 1, \ldots, H$, as recommended by Brooks and Gelman (1998). Each pair of chains is divided into batches of length $b = 10$, so that the hth value of the statistic is based in the latter half of $2hb$ observations, where $H = 150$. For each pair of chains, h_c is found as the smallest h where $\sqrt{\widehat{R}(h)}$ never exceeds 1.01. The burn-in is set to b times the largest h_c across all pairs of chains for all parameters. Graphs of $V(h)$, $W(h)$ and $\sqrt{\widehat{R}(h)}$ versus h are also inspected to make sure that $V(h)$ and $W(h)$ stabilize as a function of h and $\sqrt{\widehat{R}(h)}$ approaches 1 (Brooks and Gelman, 1998).

2.6. Posterior moments

The marginal posterior means and covariance matrix could be estimated by the sample means and sample covariance matrix of the simulated parameter values $\boldsymbol{\vartheta}^k$. However, it is more efficient to exploit the means $\widehat{\boldsymbol{\vartheta}}^k$ and variances $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta}}$ of the conditional posterior distributions. Gelfand and Smith (1990) pointed out that this follows from the Rao–Blackwell theorem. Let n be the total number of iterations and p the number of iterations to reach convergence, i.e., the burn-in. The marginal posterior means are estimated by the mean of the conditional posterior means

$$\text{E}(\widehat{\boldsymbol{\vartheta}}|\mathbf{y}) \approx \overline{\boldsymbol{\vartheta}} \equiv \frac{1}{n-p}\sum_{k=p+1}^{n}\widehat{\boldsymbol{\vartheta}}^k. \tag{17}$$

The corresponding covariance matrix is estimated by

$$\text{Var}(\widehat{\boldsymbol{\vartheta}}|\mathbf{y}) \approx \frac{1}{n-p}\sum_{k=p+1}^{n}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta}} + \frac{1}{n-p-1}\sum_{k=p+1}^{n}(\widehat{\boldsymbol{\vartheta}}^k - \overline{\boldsymbol{\vartheta}})(\widehat{\boldsymbol{\vartheta}}^k - \overline{\boldsymbol{\vartheta}})'. \tag{18}$$

The first term is the sample mean of the conditional posterior (or "within-imputation") variance and the second term is the sample variance of the conditional posterior means (or the "between-imputation" variance) (Rubin, 1987). In the context of multiple imputation of missing data, Rubin (1987) argues that large-sample relative efficiency is high enough when the number of imputations (i.e., $n - p$) is as low as 5. We use his small-sample correction which consists of multiplying the second term by $1 + 1/(n - p)$.

2.7. Algorithm comparison for salamander mating data

We consider the widely used salamander mating data for the purpose of algorithm comparison. The data comes from three experiments conducted by S. Arnold and P. Verell at the University of Chicago, Department of Ecology and Evolution. The first experiment in the summer of 1986 used two groups of 20 salamanders from two distinct populations called roughbutts and whitesides. Each group comprised 5 roughbutt males (rbm), 5 whitesides males (wsm), 5 roughbutt females (rbf), and 5 whitesides female (wsf). Within each group, 60 male–female pairs were formed so that each salamander had 3 partners from the same population and 3 partners from the other population. The response is coded 1 if salamanders mate successfully and 0 otherwise. Two further experiments were performed with the same design, the first of which used the same salamanders.

Following most of the papers in which the salamander data have been analyzed, we consider model A of Karim and Zeger (1992):

$$\text{logit}\left[\Pr(y_{ij} = 1|x_{2i}, x_{3j}, \zeta_{1i}, \zeta_{2j})\right] = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3j} + \beta_4 x_{2i}x_{3j} + \zeta_{1i} + \zeta_{2j}, \tag{19}$$

Table 1
AIP estimates (SE) for the Salamander data: discrete posterior.

	Male wing	Fem. wing	Mean
Fixed part			
β_1 [cons]	1.02(0.41)	1.02(0.41)	1.02(0.41)
β_2 [wsm]	−0.70(0.48)	−0.70(0.48)	−0.70(0.48)
β_3 [wsf]	−2.96(0.58)	−2.97(0.58)	−2.96(0.58)
β_4 [wsm × wsf]	3.64(0.65)	3.64(0.65)	3.64(0.65)
Random part			
$\sqrt{\psi_1}$ [Males]	1.11(0.26)		
$\sqrt{\psi_2}$ [Females]		1.17(0.27)	

Table 2
Approximate maximum likelihood estimates for the Salamander data.

	AQ-3pt xtmelogit		MCEM[a]	PQL[b]		Laplace				H-Likelihood[c]	Pairwise likelihood[d]	
						xtmelogit		lmer				
	Est	SE	Est	Est	SE	Est	SE	Est	SE	Est	Est	SE
Fixed part												
β_1 [cons]	1.01	0.41	1.02	0.79	0.32	1.00	0.39	1.00	0.37	1.02	1.07	0.46
β_2 [wsm]	−0.70	0.48	−0.69	−0.54	0.39	−0.70	0.46	−0.70	0.44	−0.72	−0.73	0.53
β_3 [wsf]	−2.95	0.58	−2.96	−2.29	0.43	−2.90	0.56	−2.91	0.50	−2.97	−3.09	0.65
β_4 [wsm × wsf]	3.62	0.64	3.63	2.82	0.50	3.59	0.64	3.59	0.54	3.66	3.81	0.75
Random part												
$\sqrt{\psi_1}$ [Males]	1.10		1.12	0.79		1.02		1.03		1.10	1.26	0.34
$\sqrt{\psi_2}$ [Females]	1.16		1.18	0.72		1.08		1.08		1.18	1.30	0.33

[a] Booth and Hobert (1999), Derived from Table 6 (different parameterization); Within .01 of estimates derived from Table 2 in Vaida and Meng (2005).
[b] Breslow and Clayton (1993), Tables 8 and 9.
[c] Lee et al. (2006), Table 6.3.m HL(2), SE not reported.
[d] Bellio and Varin (2005), Table 2. SE computed as range of 90% CI/3.3.

Table 3
MCMC estimates for the Salamander data.

	MCMC[a]		MCMC/Half normal			MCMC/Half Cauchy			MCMC/Uniform			MCMC/Inv-Gamma		
	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Fixed part														
β_1 [cons]	1.03	0.43	1.07	1.06	0.49	1.05	1.04	0.42	1.09	1.07	0.45	1.06	1.05	0.44
β_2 [wsm]	−0.69	0.50	−0.73	−0.71	0.54	−0.73	−0.72	0.51	−0.75	−0.74	0.52	−0.72	−0.71	0.50
β_3 [wsf]	−3.01	0.60	−3.18	−3.14	0.67	−3.06	−3.03	0.58	−3.17	−3.14	0.63	−3.06	−3.03	0.62
β_4 [wsm × wsf]	3.74	0.68	3.90	3.87	0.71	3.75	3.72	0.66	3.90	3.88	0.69	3.74	3.72	0.68
Random part														
$\sqrt{\psi_1}$ [Males]	1.17	0.28	1.28	1.27	0.30	1.17	1.16	0.28	0.83	0.80	0.21	1.19	1.18	0.28
$\sqrt{\psi_2}$ [Females]	1.22	0.29	1.35	1.32	0.31	1.24	1.22	0.28	0.78	0.75	0.22	1.24	1.22	0.29

[a] Karim and Zeger (1992), Table 3. SE computed as range of 90% CI/3.3.

where x_{2i} is a dummy variable for being a whiteside male, x_{3j} is a dummy variable for being a whiteside female, ζ_{1i} is a random intercept for male i, and ζ_{2j} is a random intercept for female j. The random intercepts are assumed to be independently distributed as $\zeta_{1i}|x_{2i}, x_{3j} \sim \mathrm{N}(0, \psi_1)$ and $\zeta_{2j}|x_{2i}, x_{3j} \sim \mathrm{N}(0, \psi_2)$. Here the salamanders from experiments 1 and 2 are treated as independent.

Ten quadrature points were used for estimation in the male and female wings. Both a normal and a discrete approximation were used for sampling from the posterior distributions of the random effects. The burn-in was found to be 340 for the normal approximation and 190 for the discrete approximation. Using the graphical method, the three statistics ($V(h)$, $W(h)$, and $\widehat{R}(h)$, $h = 1, \ldots, H$) stabilized as a function of h.

An additional 2600 iterations after burn-in were used for the estimation of posterior moments. Results for the discrete posterior are shown in Table 1. Results for the normal approximation never differed by more than 0.01 from those of Table 1.

Estimates using alternative methods are given in Tables 2 and 3. Table 2 shows results from 3-point adaptive quadrature (combined with reformulation of the model as described in the introduction) implemented in Stata's *xtmelogit* command (StataCorp, 2007), Monte Carlo EM (MCEM) (Vaida and Meng, 2005), PQL (Breslow and Clayton, 1993), Laplace implemented in Stata's *xtmelogit* command and the *R* function *lmer* (Bates et al., 2008), H-likelihood, HL(2) (Lee et al., 2006), and pairwise likelihood (Bellio and Varin, 2005). Table 3 shows MCMC results from Karim and Zeger (1992), who specified uniform priors for the variances, as well as results using WinBUGS 1.4 (Spiegelhalter et al., 2003) with four different priors: half normal on the standard deviation (mean 0, variance 10 000), half-Cauchy on the standard deviation (as illustrated in the Appendix of Gelman, 2006), locally uniform on the standard deviation (from 0 to 100), and inverse-Gamma on the variance

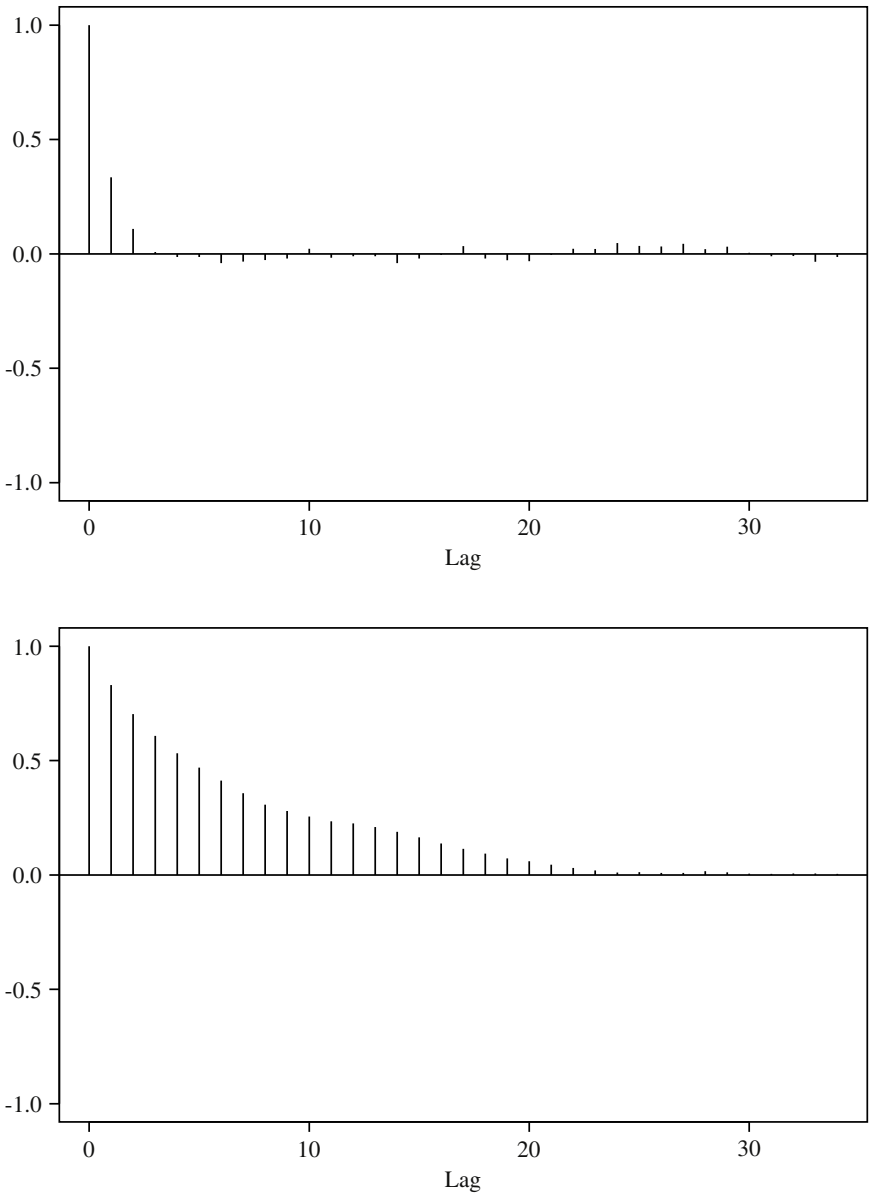

Fig. 1. Sampler lag-autocorrelation for $\sqrt{\psi_1}$ using AIP with normal approximation (top panel) and Gibbs sampler (bottom panel).

(parameters 10 000, 10 000). Convergence checking was performed in WinBUGS using similar graphical checks as described in Section 2.5 along with the condition that $\sqrt{\widehat{R}}$ is less than 1.001.

The results for the fixed effects are nearly identical across methods. The estimates of the standard deviations of the random effects are similar for AIP with adaptive quadrature, *xtmelogit*/adaptive quadrature, MCEM, and the H-likelihood method, HL(2). Considering MCEM the gold-standard because, apart from simulation noise, it corresponds to maximum likelihood estimation, the PQL and Laplace estimates implemented in *xtmelogit* and *lmer* are too low. The reason for this is probably the small cluster size of 6, with each salamander mating with 6 salamanders. The pairwise likelihood estimates are too high. The estimates from MCMC are different from those of other algorithms, and the results depend on the choice of hyperprior.

Karim and Zeger (1992) reported the medians of the posterior distributions. Their estimates (based on uniform hyperpriors for variances) are similar to the medians based on inverse-Gamma hyperpriors for the variances and the means or medians based on half-Cauchy on the standard deviation.

Sampler lag-autocorrelations for the person standard deviation are shown in Fig. 1 for the AIP algorithm (normal approximation) and the Gibbs sampler (with inverse-Gamma hyperprior distributions for the variance parameters). The autocorrelation declines considerably more rapidly with increasing lag for AIP than for the Gibbs sampler.

2.8. Algorithm comparison with three-level data

We use three-level data from Rodríguez and Goldman (1995) and Clayton and Rasbash (1999) to compare our version of AIP with adaptive quadrature to the version used by Clayton and Rasbash (1999) with PQL-2. Clayton and Rasbash (1999)

Table 4
Estimates (SE) using different methods for simulated dataset 3 from Rodríguez and Goldman (1995).

	True parameter	MCMC Gamma[a]	Adative (10pt)	AIP Adaptive		AIP PQL-2[b]	
				F-wing (12pt)	C-wing (10pt)	F-wing	C-wing
β_0	0.665	0.557	0.556	0.557	0.557	0.507	0.520
		(0.196)	(0.191)	(0.195)	(0.192)	(0.167)	(0.175)
β_1	1.000	1.125	1.121	1.123	1.128	1.041	1.037
		(0.225)	(0.224)	(0.225)	(0.226)	(0.202)	(0.206)
β_F	1.000	1.048	1.044	1.042	1.046	0.984	0.973
		(0.119)	(0.116)	(0.117)	(0.117)	(0.100)	(0.100)
β_C	1.000	0.929	0.926	0.927	0.928	0.857	0.867
		(0.253)	(0.244)	(0.249)	(0.244)	(0.214)	(0.223)
ψ_F	1.000	0.979	0.962	0.969		0.527	
		(0.322)	(0.306)	(0.310)		(0.155)	
ψ_C	1.000	0.847	0.800		0.809		0.695
		(0.194)	(0.182)		(0.185)		(0.136)

[a] Clayton and Rasbash (1999), Table 3.
[b] Clayton and Rasbash (1999), Table 5.

compared the performance of AIP combined with MQL-1, MQL-2, PQL-1, and PQL-2 using the third of 100 datasets simulated by Rodríguez and Goldman (1995). The data are binary responses generated from a three-level random intercept logistic regression model for 2249 subjects nested in 1558 families from 161 communities. Although the classifications of family and community are nested, the AIP algorithm can be applied in the same way as for crossed classifications. The model included a subject-level covariate with coefficient $\beta_1 = 1$, a family-level covariate with coefficient $\beta_F = 1$, a community-level covariate with coefficient $\beta_C = 1$, and the intercept was $\beta_0 = 0.665$. The variance components at the family and community level were $\psi_F = 1$ and $\psi_C = 1$. The estimates of ψ_F reported in Clayton and Rasbash (1999) for AIP with MQL-1, MQL-2, PQL-1, PQL-2, are, respectively 0.304, 0.393, 0.329, and 0.527. For ψ_C, the estimates are, in the same order, 0.518, 0.559, 0.598, and 0.695. It is not surprising that the estimate of the family level variance are particularly low since there are on average only 1.4 members per family, and the MQL/PQL approximations are known to perform poorly for binary responses when cluster sizes are small.

Table 4 reports the AIP with PQL-2 estimates from Clayton and Rasbash (1999) together with our AIP estimates with adaptive quadrature. For AIP with adaptive quadrature, 12 quadrature points were used for estimation in the family wing and 10 quadrature points were employed for estimation in the community wing. The burn-in was set to be 130 and an additional 1800 iterations after burn-in were used for the estimation of posterior moments. Also shown in Table 4 are MCMC estimates from Clayton and Rasbash (1999), with uniform priors on fixed effects and inverse-Gamma priors (parameters 909, 909) for the variances parameters, and maximum likelihood estimates using adaptive quadrature with 10 quadrature points per random effect using *gllamm*. There is little difference in the estimates of fixed effects across different methods except for AIP with PQL-2. While the family-level variance is severely underestimated using AIP with PQL-2, the estimate using AIP with adaptive quadrature is very close to the true value and to the estimates using MCMC or adaptive quadrature. The community-level variance is also underestimated using AIP with PQL-2 in comparison to the other three methods.

3. Item response models with random item parameters

Item response models nearly always have random person parameters (e.g., Bock and Lieberman, 1970). Due to the incidental parameter problem (Neyman and Scott, 1948), this approach is preferable to treating both items and persons as fixed and estimating item and person parameters simultaneously. It is also more appropriate to consider the person parameter as random when we are interested in making inferences regarding the population, not just the persons included in the sample.

Item response models with random item parameters are also sometimes specified (e.g., Van den Noortgate et al., 2003). Such models are appropriate if items are sampled from an item bank. One application in educational testing is sampling a different set of items from the same pool of items at different time points to prevent cheating by item exposure (Albers et al., 1989). Even for items that are not randomly sampled, we may want to generalize to the "universe of items" as in generalizability theory (see also Briggs and Wilson, 2007). Sometimes items can be computer-generated "on the fly" from an item model or item family. The item parameters are then not known for the purpose of scoring, but a model with random item parameters can be used (Janssen et al., 2000; Sinharay et al., 2003; Johnson and Sinharay, 2005; Glas and van der Linden, 2003) in which the distribution of the item parameters for the item family has been estimated. In addition, if the model includes item covariates as in explanatory item response models (De Boeck and Wilson, 2004), it is natural to allow for a random residual in the "item regression" (De Boeck, 2008; Janssen et al., 2004).

Random item difficulty parameters have also been used for modeling differential item functioning (DIF) between groups of examinees. The traditional model-based approach is to introduce a group by item interaction in an item response model. Chaimongkol et al. (2006) extend such models to a multilevel setting by including a random intercept for schools and a school-level random group by item interaction to model variability in the severity of DIF across schools. De Jong and

Table 5
Different estimates for the math test data.

	AIP[a]		Laplace xtmelogit		MCMC											
					Half normal			Half Cauchy			Uniform			Inv-Gamma		
	Est	SE	Est	SE	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Fixed part																
β_1	1.17	0.19	1.17	0.19	1.17	1.17	0.24	1.13	1.15	0.13	1.14	1.12	0.22	1.23	1.23	0.19
Random part																
$\sqrt{\psi_1}$ [Person]	1.16	0.03	1.14	0.03	1.16	1.16	0.03	1.16	1.16	0.03	0.86	0.86	0.02	1.16	1.16	0.03
$\sqrt{\psi_2}$ [Item]	0.79	0.13	0.79	0.13	0.89	0.87	0.17	0.83	0.81	0.15	1.17	1.16	0.21	0.85	0.82	0.16

[a] Identical results for normal and discrete approximations.

Steenkamp (2010) allow the thresholds and item discrimination parameters in a graded response model for polytomous items to vary randomly between countries to investigate DIF across countries. De Jong and Steenkamp (2007) extend this model to multidimensional latent traits and by specifying a finite mixture distribution for the random thresholds and discrimination parameters. De Boeck (2008) specifies a latent class model on the item side, with items belonging to one latent class displaying DIF and items belonging to the other latent class being free from DIF. For DIF items, the item difficulties for the focal and reference groups are assumed to follow a bivariate normal distribution with different means and variances for the two groups. For non-DIF items, the difficulties have a normal distribution with mean and variance the same across groups and equal to the mean and variance for the DIF items in the reference group.

3.1. Empirical study

Here we consider a dataset collected by Doolaard (1999) and previously analyzed by Fox and Glas (2001) and Vermunt (2007). The data are from an 18-item math test taken by 2156 pupils belonging to 97 schools in the Netherlands. We estimated the model described in the introduction (i.e., Eq. (1)). The dataset was chosen because it is publicly available (http://www.statisticalinnovations.com/products/latentgold_datasets.html), allowing future comparison with other methods. Further, by considering a simple model with a random intercept for each of two cross-classifications (items and persons), this example resembles examples from other disciplines such as panel data models with two-way error components, or random-intercept models for students nested in schools cross-classified by neighborhoods.

Fifteen quadrature points were used for the person wing and 5 quadrature points for the item wing. Convergence was achieved after less than 70 iterations for all parameters using the normal approximation and 150 iterations for all parameters using the discrete posterior approximation. An additional 2800 iterations were used to calculate the posterior moments after burn-in.

Table 5 shows estimates using AIP, the Laplace approximation implemented in *xtmelogit*, and MCMC using the same priors/hyper-priors as for the salamander data. The fixed effects estimates are similar across methods except for MCMC with inverse-Gamma priors on the variances. The standard deviation estimates for the person random effect are similar across methods except for MCMC with locally uniform priors on the standard deviations. However, the standard deviation estimates for the item random effect differ considerably between methods. The Bayesian estimates of the standard deviation of the item random effect are a little larger than the approximate maximum likelihood estimates, except for the MCMC estimate with locally uniform priors which is considerable larger.

4. Simulation study

In this section, we present a simulation study designed to assess the performance of the AIP algorithm with adaptive quadrature and the Laplace approximation across a range of conditions that are realistic either for item response data or longitudinal data.

A latent response formulation for the item response model with a random item parameter can be used to define intraclass correlations. Let there be a latent response y^*_{ij} so that the observed response is 1 if $y^*_{ij} > 0$ and 0 otherwise. Assuming that

$$y^*_{ij} = \beta_1 + \zeta_{1j} + \zeta_{2i} + \epsilon_{ij}, \tag{20}$$

and that the error ϵ_{ij} follows a logistic distribution, produces the model in Eq. (1) for the observed responses y_{ij}. We can now define an intraclass correlation for persons as the correlation among latent responses for the same person, conditional on the item difficulties,

$$\mathrm{Corr}(y^*_{ij}, y^*_{i'j}|\zeta_{2i}, \zeta_{2i'}) = \rho(\mathrm{P}) = \frac{\psi_1}{\psi_1 + \frac{\pi^2}{3}}. \tag{21}$$

Similarly, the intraclass correlation for items or occasions is defined as

$$\mathrm{Corr}(y^*_{ij}, y^*_{ij'}|\zeta_{1j}, \zeta_{1j'}) = \rho(\mathrm{I}) = \frac{\psi_2}{\psi_2 + \frac{\pi^2}{3}}. \tag{22}$$

Table 6
Estimates using AIP with a discrete approximation.

Con.	J	I	R_P	R_I	ρ(P)	ρ(I)	β_1				ψ_1				ψ_2			
							True	Bias	SD	M(SE)	True	Bias	SD	M(SE)	True	Bias	SD	M(SE)
1	100	10	5	5	0.1	0.1	0	−0.006	0.207	0.199	0.366	−0.007	0.135	0.129	0.366	−0.045*	0.181	0.168
2	100	10	5	10	0.1	0.5	0	−0.073	0.564	0.545	0.366	0.016	0.184	0.166	3.290	−0.348*	1.507	1.436
3	100	10	5	5	0.5	0.1	0	−0.030	0.258	0.274	3.290	0.049	0.684	0.718	0.366	−0.049*	0.188	0.175
4	100	10	15	10	0.5	0.5	0	−0.071	0.577	0.567	3.290	0.057	0.757	0.745	3.290	−0.426*	1.463	1.373
5	100	50	10	10	0.1	0.1	0	0.016	0.118	0.110	0.366	0.007	0.064	0.068	0.366	−0.007	0.090	0.083
6	100	50	5	15	0.1	0.5	0	0.036	0.284	0.257	0.366	0.003	0.071	0.073	3.290	−0.092	0.697	0.693
7	100	50	10	5	0.5	0.1	0	0.012	0.223	0.209	3.290	0.051	0.511	0.542	0.366	−0.007	0.091	0.087
8	100	50	10	10	0.5	0.5	0	0.032	0.357	0.326	3.290	0.042	0.520	0.532	3.290	−0.085	0.715	0.685
9	1000	10	10	5	0.1	0.1	0	0.027	0.183	0.176	0.366	0.002	0.044	0.041	0.366	−0.047*	0.160	0.145
10	1000	10	5	5	0.1	0.5	0	0.004	0.604	0.507	0.366	0.005	0.054	0.051	3.290	−0.474*	1.282	1.270
11	1000	10	15	5	0.5	0.1	0	−0.006	0.205	0.185	3.290	0.032	0.248	0.228	0.366	−0.056*	0.145	0.142
12	1000	10	15	5	0.5	0.5	0	−0.015	0.589	0.501	3.290	0.000	0.235	0.233	3.290	−0.523*	1.272	1.246
13	1000	50	10	5	0.1	0.1	0	−0.002	0.081	0.084	0.366	0.002	0.020	0.021	0.366	−0.016*	0.072	0.071
14	1000	50	5	5	0.1	0.5	0	−0.020	0.251	0.256	0.366	0.000	0.021	0.023	3.290	−0.148*	0.618	0.636
15	1000	50	15	5	0.5	0.1	0	0.003	0.095	0.102	3.290	0.008	0.153	0.169	0.366	−0.016*	0.070	0.072
16	1000	50	10	5	0.5	0.5	0	−0.015	0.243	0.264	3.290	0.023	0.159	0.167	3.290	−0.139*	0.633	0.635

* True value outside approximate 95% confidence interval.

Table 7
Estimates using the Laplace approximation implemented in *xtmelogit*.

Con.	J	I	ρ(P)	ρ(I)	β_1				ψ_1				ψ_2			
					True	Bias	SD	M(SE)	True	Bias	SD	M(SE)	True	Bias	SD	M(SE)
1	100	10	0.1	0.1	0	−0.010	0.207	0.195	0.366	−0.024	0.131	0.125	0.366	−0.046*	0.180	0.167
2	100	10	0.1	0.5	0	−0.062	0.558	0.533	0.366	−0.015	0.167	0.155	3.290	−0.408*	1.462	1.401
3	100	10	0.5	0.1	0	−0.022	0.261	0.267	3.290	−0.065	0.656	0.695	0.366	−0.039*	0.192	0.181
4	100	10	0.5	0.5	0	−0.067	0.571	0.559	3.290	−0.033	0.719	0.724	3.290	−0.436*	1.436	1.364
5	100	50	0.1	0.1	0	0.017	0.119	0.109	0.366	0.007	0.064	0.068	0.366	−0.006	0.091	0.083
6	100	50	0.1	0.5	0	0.040	0.278	0.262	0.366	0.004	0.071	0.073	3.290	−0.100	0.691	0.690
7	100	50	0.5	0.1	0	0.015	0.219	0.206	3.290	0.035	0.501	0.537	0.366	−0.002	0.093	0.088
8	100	50	0.5	0.5	0	0.032	0.348	0.315	3.290	0.053	0.514	0.533	3.290	−0.069	0.722	0.688
9	1000	10	0.1	0.1	0	0.030	0.183	0.175	0.366	−0.020*	0.043	0.040	0.366	−0.050*	0.158	0.144
10	1000	10	0.1	0.5	0	0.008	0.593	0.516	0.366	−0.028*	0.049	0.048	3.290	−0.510*	1.265	1.254
11	1000	10	0.5	0.1	0	−0.003	0.202	0.182	3.290	−0.118*	0.231	0.216	0.366	−0.060*	0.141	0.140
12	1000	10	0.5	0.5	0	0.006	0.586	0.515	3.290	−0.144*	0.217	0.220	3.290	−0.556*	1.231	1.231
13	1000	50	0.1	0.1	0	−0.004	0.081	0.086	0.366	0.000	0.020	0.021	0.366	−0.016*	0.072	0.071
14	1000	50	0.1	0.5	0	−0.014	0.237	0.251	0.366	−0.003	0.022	0.023	3.290	−0.131*	0.633	0.639
15	1000	50	0.5	0.1	0	0.000	0.096	0.102	3.290	−0.010	0.151	0.167	0.366	−0.014	0.070	0.072
16	1000	50	0.5	0.5	0	−0.012	0.241	0.257	3.290	0.011	0.158	0.166	3.290	−0.139*	0.630	0.634

* True value outside approximate 95% confidence interval.

We examined all combinations of two test lengths or numbers of occasions (10 and 50), two sample sizes (100 persons and 1000 persons), two intraclass correlations for persons (0.1 and 0.5) and two intraclass correlations for items (0.1 and 0.5), resulting in 16 conditions. For each condition, we simulated 100 datasets and estimated the model using AIP with a normal and discrete approximation and using Laplace (as implemented in Stata's *xtmelogit* command). For each condition, the same 100 datasets were analyzed by the three methods to enable accurate comparisons.

Five to 15 quadrature points were used. One simulated data set for each condition was used for convergence checking and the same burn-in was set across replications. Using the convergence check described in Section 2.5 (with $b = 10$), convergence was achieved in 10–80 iterations for the normal approximation and in 10–90 iterations for the discrete approximation. An additional 10 iterations were obtained to estimate the posterior moments.

Table 6 shows the number of quadrature points, estimated bias, standard deviation (SD) of the estimates, and mean of the standard errors (M(SE)) for the AIP algorithm with a discrete approximation. (The results for AIP with the normal and discrete approximations were very similar across simulation conditions.) Table 7 presents the results of the Laplace approximation implemented in *xtmelogit*. Asterisks indicate that there was significant bias at the 5% level using a one-sample t-test. There was no significant bias for β_1 for any of the conditions using any of the methods. There was also no significant bias for the person variance ψ_1 using the AIP algorithm.

However, using the Laplace approximation, there was significant downward bias for ψ_1 for conditions 9–12 with $J = 1000$ and $I = 10$. The estimated bias was larger when the person variance was larger. Such downward bias has previously been found for Laplace (Joe, 2008) and related methods (MQL and PQL) (Browne and Draper, 2006) for binary responses with small cluster sizes (here $I = 10$) and high intraclass correlations. It might therefore be expected that there

Table 8
Additional conditions with large person variance.

Method	J	I	ρ(P)	ρ(I)	β_1				ψ_1				ψ_2			
					True	Bias	SD	M(SE)	True	Bias	SD	M(SE)	True	Bias	SD	M(SE)
AIP/Normal	100	10	0.8	0.5	0	−0.001	0.749	0.676	13.146	−0.046	3.326	3.099	3.290	−0.465*	1.476	1.380
AIP/Discrete	100	10	0.8	0.5	0	0.013	0.696	0.666	13.146	−0.315	2.908	2.992	3.290	−0.504*	1.393	1.352
Laplace	100	10	0.8	0.5	0	−0.001	0.722	0.643	13.146	−1.074*	2.873	2.770	3.290	−0.493*	1.409	1.360
AIP/Normal	100	50	0.8	0.5	0	0.067	0.422	0.463	13.146	0.319	2.624	2.334	3.290	−0.093	0.717	0.691
AIP/Discrete	100	50	0.8	0.5	0	0.052	0.424	0.462	13.146	0.236	2.464	2.312	3.290	−0.079	0.749	0.690
xtmelogit	100	50	0.8	0.5	0	0.056	0.430	0.456	13.146	−0.045	2.326	2.239	3.290	−0.033	0.748	0.704

* True value outside approximate 95% confidence interval.

should be significant downward bias also for $I = 10$ combined with $J = 100$ (conditions 1–4), but the power is reduced due to the smaller sample size as reflected by the larger standard errors.

All methods show downward bias for the item variance ψ_2 when there are $I = 10$ items (conditions 1–4 and 9–12), with larger estimated bias for larger item variances. This downward bias may be due to the usual downward bias found in maximum likelihood estimation with a small number of clusters (e.g., Raudenbush and Bryk, 2002, p. 283). In linear models, this problem is addressed by using restricted maximum likelihood (REML) estimation (Patterson and Thompson, 1971). Unfortunately, the REML concept cannot be directly applied to generalized linear mixed models, although there are some ad-hoc approaches (Schall, 1991; Breslow and Clayton, 1993; McGilchrist, 1994; Stiratelli et al., 1984; Noh and Lee, 2007).

The assumption that downward bias is due to the small number of clusters was confirmed by eliminating the person random effect, simulating data for 10 and 100 items (with $J = 100$ and $\psi_2 = 3.290$) and finding that there is a significant downward bias for the item variance with 10 items but not with 100 items, estimated, respectively, as −0.25 (SE = 0.05) and −0.01 (SE = 0.02) using 1000 replications. Returning to Tables 6 and 7, there was also significant, but considerably less bias for the item variance for $I = 50$ combined with $J = 1000$ (conditions 13–16). The estimated biases were not considerably larger than for $I = 50$ and $J = 100$, but significance was reached in part because of smaller standard errors due to the larger sample size. The means of the estimated standard errors are quite close to the empirical standard deviations of the estimates for all parameters and conditions.

Two additional conditions were considered as shown in Table 8. The conditions are the same as conditions 4 and 8 in Tables 6 and 7 except that the person variance is now $\psi_1 = 13.146$, producing a larger $\rho(P)$ of 0.8 instead of 0.5. When $I = 10$, this larger variance resulted in a larger and significant downward bias for ψ_1 using Laplace (c.f. condition 4 in Table 7). However, the downward bias for ψ_1 was non-significant for $I = 50$. Using all three methods, the estimated bias for the item variance ψ_2 was similar to the bias estimated when the person variance was lower (c.f. conditions 4 and 8 in Tables 6 and 7).

The results suggest that estimation of a given variance component is affected mostly by the true value of that variance component, the relevant cluster size (number of items for person variance and number of persons for item variance) and the relevant number of clusters (number of persons for person variance and number of items for item variance). A small number of clusters leads to downward bias for all three methods, particularly when the corresponding true variance is large. A small cluster size is only a problem for the Laplace approximation, particularly when the corresponding cluster variance is large.

5. Discussion

The AIP algorithm gave similar results to alternative approximate maximum likelihood methods for two real datasets. In the salamander data, it performed better than PQL and Laplace, probably because of the small clusters of size 6. In the simulation study, the AIP algorithm performed well in a wide range of conditions. An exception was estimation of the random intercept variance when the corresponding number of clusters was 10, but the downward bias appears to be due to using approximate maximum likelihood estimation with a small number of clusters. The Laplace approximation performed similarly well, except for the random intercept variance when the corresponding cluster size was 10.

A disadvantage of AIP, shared by other MCMC methods, is that it can be difficult to assess convergence. However, unlike other MCMC methods, AIP does not require specification of prior distributions for model parameters. This may be an advantage since, as we have shown, the choice of hyperprior for the variance components can affect the parameter estimates. Unfortunately, AIP does not share the advantages of full Bayesian MCMC estimation which delivers the entire posterior distribution of each parameter including the random effect and does not rely on asymptotic theory (in sample size). Although not usually done, it would generally be straightforward to increase efficiency in MCMC estimation of posterior means and standard deviations by using Rao–Blackwellization.

An important advantage of the AIP algorithm is that it is easy to implement. The algorithm can be used to estimate more complex random effects and latent variable models than considered here. All that is required is that the random part of the model can be split into two or more parts, corresponding to wings in the algorithm, so that the parameters of each part can be estimated when the other random effects are held constant. A possible application in psychometrics would be two-parameter item response models with random item difficulty and discrimination parameters.

References

Albers, W., Does, R.J.M.M., Imbos, T., Janssen, M.P.E., 1989. A stochastic growth model applied to repeated test of academic knowledge. Psychometrika 54, 451–466.

Albert, P.S., Follmann, D.A., 2000. Modeling repeated count data subject to informative dropout. Biometrics 56, 667–677.

Baltagi, B.H., 2001. Econometric Analysis of Panel Data, second ed. Wiley, Chichester.

Bates, D., Maechler, M., Dai, B., 2008. The lme4: linear mixed-effects models using S4 classes. URL: http://lme4.r-forge.r-project.org/.

Bellio, R., Varin, C., 2005. A pairwise likelihood approach to generalized linear models with crossed random effects. Statistical Modelling 5, 217–227.

Bock, R.D., Lieberman, M., 1970. Fitting a response model for *n* dichotomously scored items. Psychometrika 33, 179–197.

Bock, R.D., Schilling, S.G., 1997. High-dimensional full-information item factor analysis. In: Berkane, M. (Ed.), Latent Variable Modelling and Applications to Causality. Springer, New York, pp. 164–176.

Booth, J.G., Hobert, J.P., 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. Journal of the Royal Statistical Society, Series B 61, 265–285.

Breslow, N.E., 2004. Whither PQL? In: Lin, D.Y., Heagerty, P.J. (Eds.), Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data. Springer, New York, pp. 1–22.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.

Breslow, N.E., Lin, X., 1995. Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika 82, 81–91.

Briggs, D.C., Wilson, M., 2007. Generalizability in item response modeling. Journal of Educational Measurement 44, 131–155.

Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7, 434–455.

Browne, W.J., Draper, D., 2006. A comparison of Bayesian and likelihood methods for fitting multilevel models. Bayesian Analysis 1, 473–514.

Browne, W.J., Goldstein, H., Rasbash, J., 2001. Multiple membership multiple classification (MMMC) models. Statistical Modelling 1, 103–124.

Butler, J.S., Moffitt, R., 1982. A computationally efficient quadrature procedure for the one-factor multinomial probit model. Econometrica 50, 761–764.

Chaimongkol, S., Huffer, F.W., Kamata, A., 2006. A Bayesian approach for fitting a random effect differential item functioning across group units. Thailand Statistician 4, 27–41.

Chang, H., Stout, W., 1993. The asymptotic posterior normality of the latent trait in an IRT model. Psychometrika 58, 37–52.

Clayton, D.G., Rasbash, J., 1999. Estimation in large crossed random-effect models by data augmentation. Journal of the Royal Statistical Society, Series A 162, 425–436.

De Boeck, P., 2008. Random item IRT models. Psychometrika 73, 533–559.

De Boeck, P., Wilson, M., 2004. Explanatory Item Response Models: a Generalized Linear and Nonlinear Approach. Springer, New York.

De Jong, M.G., Steenkamp, J., 2007. Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. Journal of Consumer Research 34, 260–278.

De Jong, M.G., Steenkamp, J., 2010. Finite mixture multilevel multidimensional ordinal IRT models for large-scale cross-cultural research. Psychometrika 75, 3–32.

Diaz, R.E., 2007. Comparison of PQL and Laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomised trials. Computational Statistics and Data Analysis 51, 2871–2888.

Doolaard, S., 1999. Schools in Change or School in Chain. University of Twente, The Netherlands.

Fox, J.P., Glas, C.A.W., 2001. Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika 66, 269–286.

Gelfand, A., Smith, A., 1990. Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association 90, 398–409.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1, 515–533.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statistical Sceience 7, 457–472.

Glas, C.A.W., van der Linden, W.J., 2003. Computerized adaptive testing with item cloning. Applied Psychological Measurement 27, 247–261.

Goldstein, H., 1987. Multilevel covariance component models. Biometrika 74, 430–431.

Goldstein, H., 1991. Nonlinear multilevel models, with an application to discrete response data. Biometrika 78, 45–51.

Goldstein, H., 2003. Multilevel Statistical Models, third ed. Arnold, London.

Goldstein, H., Rasbash, J., 1996. Improved approximations for multilevel models with binary responses. Journal of the Royal Statistical Society, Series A 159, 505–513.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W.J., Yang, M., Woodhouse, G., Healy, M., 1998. A User's Guide to MLwiN. Multilevel Models Project, Institute of Education, University of London, London.

Hedeker, D., Gibbons, R.D., 1994. A random-effects ordinal regression model for multilevel analysis. Biometrics 50, 933–944.

Janssen, R., Schepers, J., Peres, D., 2004. Models with item and item group predictors. In: De Boeck, P., Wilson, M. (Eds.), Explanatory Item Response Models. Springer, New York, pp. 198–212.

Janssen, R., Tuerlinckx, F., Meulders, M., De Boeck, P., 2000. A hierarchical IRT model for criterion-referenced measurement. Journal of Educational and Behavioral Statistics 25, 285–306.

Joe, H., 2008. Accuracy of Laplace approximation for discrete response mixed models. Computational Statistics and Data Analysis 52, 5066–5074.

Johnson, M.S., Sinharay, S., 2005. Calibration of polytomous item families using Bayesian hierarchical modeling. Applied Psychological Measurement 29, 369–400.

Karim, M.R., Zeger, S.L., 1992. Generalized linear models with random effects: Salamander mating revisited. Biometrics 48, 631–644.

Lambert, P.C., 2006. Comment on article by Browne and Draper. Bayesian Analysis 1, 543–546.

Lee, Y., Nelder, J.A., 1996. Hierarchical generalized linear models (with discussion). Journal of the Royal Statistical Society, Series B 58, 619–678.

Lee, Y., Nelder, J.A., 2006. Double-hierarchical generalized linear models (with discussion). Journal of the Royal Statistical Society, Series C 55, 1–29.

Lee, Y., Nelder, J.A., Pawitan, Y., 2006. Generalized Linear Models With Random Effects: Unified Analysis Via H-Likelihood. Chapman & Hall/CRC, Boca Raton, FL.

Lesaffre, E., Spiessens, B., 2001. On the effect of the number of quadrature points in a logistic random-effects model: an example. Journal of the Royal Statistical Society, Series C 50, 325–335.

Lin, X., Breslow, N.E., 1996. Bias correction in generalized linear mixed models with multiple components of dispersion. Journal of the American Statistical Association 91, 1007–1016.

Liu, Q., Pierce, D.A., 1994. A note on Gauss–Hermite quadrature. Biometrika 81, 624–629.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., 2004. Models for value-added modeling of teacher effects. Journal of Educational and Behavioral Statistics 29, 67–101.

McCulloch, C.E., 1994. Maximum likelihood variance components estimation for binary data. Journal of the American Statistical Association 89, 330–335.

McGilchrist, C.A., 1994. Estimation in generalized mixed models. Journal of the Royal Statistical Society, Series B 56, 61–69.

Mislevy, R.J., 1991. Randomization-based inference about latent variables from complex samples. Psychometrika 56, 177–196.

Mislevy, R.J., Beaton, A.E., Kaplan, B., Sheehan, K.M., 1992. Estimating population characteristic from sparse matrix samples of item responses. Journal of Educational Measurement 29, 133–161.

Natarajan, R., Kass, R.E., 2000. Reference Bayesian methods for generalized linear mixed model. Journal of the American Statistical Association 95, 227–237.

Naylor, J.C., Smith, A.F.M., 1982. Applications of a method for the efficient computation of posterior distributions. Journal of the Royal Statistical Society, Series C 31, 214–225.

Neyman, J., Scott, E.L., 1948. Consistent estimates based on partially consistent observation. Econometrica 16, 1–32.

Noh, M., Lee, Y., 2007. REML estimation for binary data in GLMMs. Journal of Multivariate Analysis 98, 896–915.

Patterson, H.D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545–554.
Pinheiro, J.C., Bates, D.M., 1995. Approximation to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphics and Statistics 4, 12–35.
Pinheiro, J.C., Chao, E.C., 2006. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. Journal of Computational and Graphical Statistics 15, 58–81.
Rabe-Hesketh, S., Skrondal, A., 2008. Multilevel and Longitudinal Modeling Using Stata, second ed. Stata Press, College Station, TX.
Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2005. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. Journal of Econometrics 128, 301–323.
Rasbash, J., Browne, W.J., 2007. Non-hierarchical multilevel models. In: de Leeuw, J., Meijer, E. (Eds.), Handbook of Multilevel Analysis. Springer, New York, pp. 333–336.
Rasbash, J., Goldstein, H., 1994. Efficient analysis of mixed hierarchical and crossed random structures using a multilevel model. Journal of Educational Statistics 337–350.
Raudenbush, S.W., 1993. A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. Journal of Educational and Behavioral Statistics 18, 321–349.
Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models. Sage, Thousand Oaks, CA.
Raudenbush, S.W., Yang, M., Yosef, M., 2000. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. Journal of Computational and Graphical Statistics 9, 141–157.
Rodríguez, G., Goldman, N., 1995. An assessment of estimation procedures for multilevel models with binary responses. Journal of the Royal Statistical Society, Series A 158, 73–89.
Rodríguez, G., Goldman, N., 2001. Improved estimation procedures for multilevel models with binary response: a case study. Journal of the Royal Statistical Society, Series A 164, 339–355.
Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.
Schall, R., 1991. Estimation in generalized linear-models with random effects. Biometrika 78, 719–727.
Schilling, S., Bock, R.D., 2005. High dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. Psychometrika 70, 533–555.
Sinharay, S., Johnson, M.S., Williamson, D.M., 2003. Calibrating item families and summerizing the results using family expected response functions. Journal of Educational and Behavioral Statistics 28, 295–313.
Spiegelhalter, D., Thomas, A., Best, N., 2003. WinBUGS version 1.4 [Computer program]. UK: MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.
StataCorp,, 2007. Statistical Software: Release 10.0. [Computer program]. College Station, TX, Stata Corporation.
Stiratelli, R., Laird, N.M., Ware, J.H., 1984. Random effects models for serial observations with binary responses. Biometrics 40, 961–971.
Tanner, M.A., 1996. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. Springer, New York.
Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–540.
Thomas, N., Gan, N., 1997. Generating multiple imputations for matrix sampling data analyzed with item response models. Journal of Educational and Behavioral Statistics 22, 425–445.
Tibaldi, F.S., Verbeke, G., Molenberghs, G., Renard, D., Van den Noortgate, W., De Boeck, P., 2007. Conditional mixed models with crossed random effects. British Journal of Mathematical and Statistical Psychology 60, 351–365.
Tierney, L., Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 81, 82–86.
Vaida, F., Meng, X.L., 2005. Two slice-EM algorithms for fitting generalized linear mixed models with binary response. Statistical Modelling 5, 229–242.
Van den Noortgate, W., De Boeck, P., Meulders, M., 2003. Cross-classification multilevel logistic models in psychometrics. Journal of Educational and Behavioral Statistics 28, 369–386.
Vermunt, J., 2007. Multilevel mixture item response theory models: an application in education testing. Bulletin of the International Statistical Institute 1–4.
Wei, G.C.G., Tanner, M.A., 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association 85, 699–704.